STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21 (34) 2010

**Robert Milewski**
**Paweł Malinowski**
**Anna Justyna Milewska**
**Piotr Ziniewicz**
Department of Statistics and Medical Informatics,
Medical University of Bialystok

**Sławomir Wołczyński**
Department of Reproduction
and Gynecological Endocrinology,
Medical University of Bialystok

# THE USAGE OF MARGIN-BASED FEATURE SELECTION ALGORITHM IN IVF ICSI/ET DATA ANALYSIS

**Abstract**: In the case of infertility treatment, successful classification will facilitate understanding of various factors affecting the success of the process. Classification itself is an important data mining problem. Many classifications and constructions of the classifier algorithms are not able to cope with the analysis of the huge amount of factors associated with this process. Feature selection allows to significantly reduce the volume of analyzed data, while maintaining the classifier prediction quality. This leads to the rejection of nonessential measurements and time reductions.

**Keywords**: IVF ICSI/ET, infertility treatment, data mining, feature selection, margin, nearest neighbor classifier

## Introduction

Infertility treatment is a process whose effectiveness depends on many different factors [1]. Specially designed and developed system to collect data of patients treated for infertility using the IVF ICSI/ET method was introduced at the Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok [2]. Although conducted research in recent years and some successful treatment predictors identification – all factors which significantly affect the final treatment results – still cannot be fully listed. Also, the prediction for the final result is not likely to be made possible. Therefore, research teams analyzing the infertility treatment effectiveness refer to the increasingly sophisticated statistical methods and bioinformatics. Some results in the prediction of treatment failure were ob-

ISBN 978–83–7431–266–0 ISSN 0860-150X 

tained after the neural networks application [3]. These results offer hope to find more advanced methods for predicting treatment effectiveness and convincing us to further utilization of the possibilities opened up by bioinformatics.

The issue of observation classification is one of the most important problems in data mining. In the context of supervised classification, it is to draw a conclusions from data based on gathered earlier results. In the process of the so-called conclusion-drawing (classifier building) chosen algorithm analyzes this data to search patterns. As a result of this a process set of rules (classifier) is created, which allow proper label assignment for a new observation. A good example is the patient's prognosis analysis relying on identified symptoms. Good classifier should be primarily characterized by high speed of execution and correctness of results, also for new observation (so called bias). Intuitively, this imposes a requirement of its intrinsic simplicity – fewer (possibly biased) rules make smaller computational costs and result in lower bias of the entire classifier. Simplicity of the final classifier makes it easier to interpret. It is a classic realization of Ockham's razor.

One of the major problems of effective conclusion-making process is the curse of the dimensionality phenomenon. Input data can have huge number of features (large dimensionality), but only some of them are significant in the classification process. Excessive features obscure regular patterns in the data, decreasing the signal-to-noise ratio. Time cost of conclusion-drawing and the target classification can grow very rapidly with the increasing dimensionality of the input. A very good example of multidimensional data are medical data, which usually contain a lot more features than observations (the ratio seeking up to $10^5$ for the gene expression analysis). To cope with the curse of dimensionality, various techniques exist, such as "regularization" (boosting [4], bagging [5]), kernel methods [6], and others. One of them is feature selection and extraction, which represents dimensionality reduction approach.

## Feature selection

Feature selection is a process of choosing proper feature subset from all such possible subsets. Feature selection is close to the feature extraction task, often these terms are used interchangeably in literature. In this article it is assumed, that feature extraction consists of the feature construction and feature extraction phase. Feature construction is a process of linking and transforming low level features into higher one. An example of such

technique is the PCA [7], or picture conversion from color to grayscale. This article will focus only on the feature selection phase, assuming that high level features are already constructed.

As a result of the feature selection process a certain feature subset is chosen, which satisfies certain criteria, captures relevant properties of the data, and is also useful in the context of used classifier. Commonly used concepts of relevance and usefulness of a feature subset were established on set theory and probabilistic ground. Feature subset $c' \subseteq C$ is called optimal (1), when accuracy $acc(*)$ of $W_{cl}$ classifier for $\boldsymbol{X}/c'$ set (dataset formed the basis of $\boldsymbol{X}$, in which all data related to features not belonging to subset $c'$ where removed) will be maximum among all such $\boldsymbol{X}/c$ subsets. The feature is called useful, when it belongs to optimal subset.

$$\forall_{c \subseteq C}\ acc(W_{cl}(\boldsymbol{X}/c')) \leq acc(W_{cl}(\boldsymbol{X}/c)) \tag{1}$$

$$\forall_{c \subseteq C-\{F_i\}}\ P(K|c \cup \{F_i\}) = P(K|c) \tag{2}$$

$$P(K|C) = P(K|C - \{F_i\}) \wedge \exists_{c \subset C-\{F_i\}}\ P(K|c \cup \{F_i\}) \neq P(K|c) \tag{3}$$

$$P(K|C) \neq P(K|C - \{F_i\}) \tag{4}$$

$$\exists_{c \subset C-\{F_i\}}\ P(C - c - \{F_i\}, K|c \cup \{F_i\}) = P(C - c - \{F_i\}, K|c) \tag{5}$$

where: $c'$ optimal feature subset; $\boldsymbol{X}$ data set; $C$ set of all features from $\boldsymbol{X}$ data set; $c$ feature subset; $W_{cl}$ classifier related to final conclusion-making; $acc(*)$ classifier accuracy; $F_i$ feature; $P(*)$ probability; $K$ class distribution.

In terms of relevance, features were divided into redundant (2), weak relevant (3) and strong relevant (4). All probabilities listed in (2)–(4) formulas are conditional probabilities of class distribution against certain feature subset. Presence of strong relevant features changes such distribution (4). Presence of irrelevant features does not affect this distribution (2). In case of the weak relevant feature, certain subsets of features exist for which presence of this feature changes the conditional distribution (3).

Depending on other features, a weak relevant one can be locally strong relevant, locally irrelevant. Taking into account local feature relevance phenomena, feature set can be divided in 4 subsets. Figure 1 presents a review of those subsets (1 – strong relevant feature subset, 2 – weak relevant, but locally relevant feature subset, 3 – weak relevant, but locally irrelevant feature subset, 4 – irrelevant feature subset). Optimal feature subset should contain all strong relevant features and weak relevant, but locally relevant ones (feature subsets 1+2 from Figure 1). Collection of irrelevant and weak

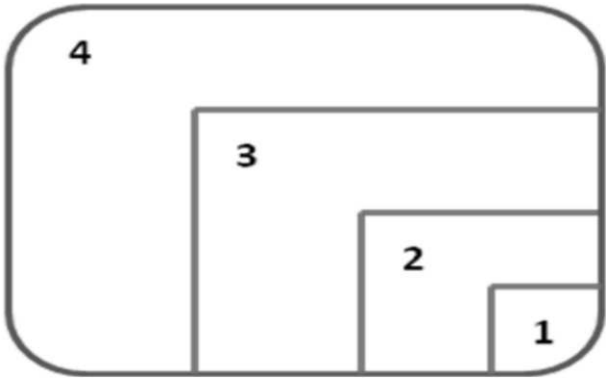


**Fig. 1. Set of features division according to their relevancy**

relevant, but locally irrelevant features make redundant the feature set (feature subsets 3+4 from Figure 1). Feature $F_i$ is redundant, if equation (5) is satisfied. All these definitions were presented in [8].

The selection of a relevant feature can be accomplished in two ways: by feature ranking or subset selection. In case of feature ranking, score is assigned for each feature, according to certain criterion. The next step are rejection features below chosen score threshold. In the subset selection approach, each possible subset of features is scored by criterion function according to three different models: filter, wrapper or embedded.

In the filter model, space of feature subsets is searched, and in each search step simple filter is used to measure subset score. This filter is usually a statistical measure, or is based on entropy, or the heuristic approach. In a special case, when single feature is treated as a subset, filter model reduces to earlier mentioned feature ranking process. In the wrapper model, space of feature subsets is also searched but in each step of searching final learning algorithm is launched. Based on its results, feature subsets comparison is possible. Embedded models include a group of algorithms, which are characteristic for a process of conclusion-drawing. These are similar to the wrapper models, the difference is, that the process of conclusion-drawing itself directs search and evaluation of different subsets of features.

## Margin-based feature selection algorithm

Margin is a geometrical measure of certainty and generalization abilities given by the classifier. Many conclusion-making algorithms and classifiers (e.g. SVM [6]) make use of the margin concept. Articles [9, 10] proposed

margin as a quality measure of feature subset that generate this margin. When searching for an optimal feature subset the algorithm tends to increase such margin. This article suggests the SIMBAF algorithm, which is a modification of the SIMBA [10] algorithm. It is a subset selection method type implementing the filter model. Algorithm accepts data which features:

1. numeric – taking values from real number field
2. ordinal – features which values equivalent to natural numbers (by article convention with 0), order is maintained, not necessarily mutual distance. For such features the distance $d$ between 0 and 1 values not necessarily the same as f. e. distance between 1 and 2 values, but condition $d(0,1) < d(0,2)$ always occurs, and "1" value is between "0" and "2" values.
3. categorical – features which values equivalent to natural numbers (by article convention with 0), order is not maintained, neither mutual distance.

$$\Delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{6}$$

$$\Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\{ \sum_i \left[ \alpha(t_i) \phi_{num}(x_{1i}, x_{2i}) \right]^p \right\}^{1/p} \tag{7}$$

$$\Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_j \alpha(t_j) \phi_{cat}(x_{1j}, x_{2j}) \tag{8}$$

$$\Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_k \alpha(t_k) \phi_{ord}(x_{1k}, x_{2k}) \tag{9}$$

where: $\boldsymbol{x}_1, \boldsymbol{x}_2$ observation; $\Delta(*,*)$ observation distance measure, low index means parts: numerical $\Delta_{num,p}(*,*)$, categorical $\Delta_{cat}(*,*)$, ordinal $\Delta_{ord}(*,*)$; $i, j, k$ indexes; $p$ metric order; $t_i, t_j, t_k$ "hidden" parameters, $\alpha(*)$ – specific function, $x_{1*}, x_{2*}$ value of feature $*$ for given observation; $\phi_*(*,*)$ measures of difference between single feature values.

Derivation of target margin form requires definition of the distance between compared observations. It was defined according to (6). Proper parts of expression (6) are related to the following features: numerical (7), categorical (8) and ordinal (9). Indexes $i, j, k$ iterate respectively over selected features types. Introduced later index $o$ will iterate over all features. Proposed algorithm breaks with using modified Euclidean metrics for numerical features [10], replacing it with the Minkowski metric of any order $p$ (7). It should be noted, that for $p < 1$, related formula (7) is not formally metric anymore, because triangle inequity axiom is not satisfied (for such cases opposite direction inequality occurs). As it will be later discussed, change of this parameter leads to interesting and important conclusions and gene-

ralizations. For categorical and ordinal features, due to their nature, metric cannot be introduced as such, and therefore special functions of the similarity are introduced to replace it.

$$\alpha(t_o) = \frac{1}{\pi}\left(\text{arctg}(t_o) + \frac{\pi}{2}\right) = \frac{\text{arctg}(t_o)}{\pi} + \frac{1}{2} \qquad t_0 \in \Re; \quad \forall_{t_o}\ \alpha(t_o) \in (0,1)... \tag{10}$$

$$\delta_o = \max_{b,d} |x_{bo} - x_{do}|; \quad 0 \le \varepsilon_{o1} \le \varepsilon_{o2} \le 1 \tag{11}$$

$$\phi_{num}(x_{1i}, x_{2i}) = \min\left(1, \max\left(0, \frac{|x_{1i} - x_{2i}| - \delta_i \varepsilon_{i1}}{\delta_i(\varepsilon_{i2} - \varepsilon_{i1})}\right)\right) \in [0,1] \tag{12}$$

$$\phi_{cat}(x_{1j}, x_{2j}) = \begin{cases} 0 & \iff x_{1j} = x_{2j} \\ 1 & \iff x_{1j} \neq x_{2j} \end{cases} \tag{13}$$

$$\phi_{ord}(x_{1k}, x_{2k}) = \min\left(1, \max\left(0, \frac{|x_{1k} - x_{2k}| - \delta_k \varepsilon_{k1}}{\delta_k(\varepsilon_{k2} - \varepsilon_{k1})}\right)\right) \in [0,1] \tag{14}$$

where: $\delta_o$ value range of $o$-th feature; $\varepsilon_{o1}$ i $\varepsilon_{o2}$ cut-off parameters for $o$-th feature (there are only 2 such parameters)

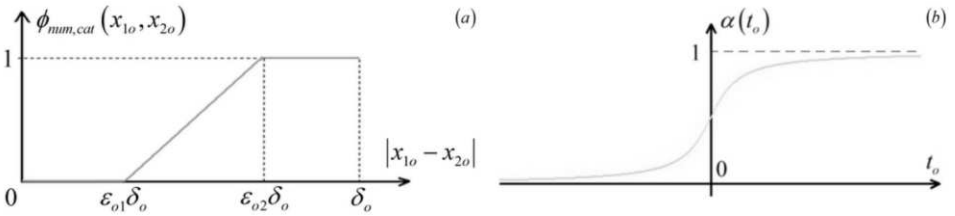


**Fig. 2. (a) distance measure for numerical and order features (b) factor $\alpha(t_o)$**

Factor $\alpha(t_o)$ plays an important role, it is a scale factor and also calculated weight of $o$-th feature. In proposed algorithm it is a function (10) of optimized "hidden" parameter $t_o$ (it is used internally). Figure 2b presents the plot of this function. The use of proper function allows to regulate calculated weights. SIMBAF algorithm is not limited to this single function, actually any other sigmoid function normalized to (0,1) can be applied. Functions $\phi(*,*)$ determine the measure of distance between values of a given feature compared observation. They were identified separately for the numerical features (12), categorical (13) and ordinal (14). All these distances were also normalized to the range [0,1] in order to normalize their impact on the target margin. For numeric and order features, the so-called

cut-off parameters $\varepsilon_{o1}$ and $\varepsilon_{o2}$ are introduced, which are designed to filter out extreme values directly modifying the distance function. They were taken from article [11]. The behavior of measurement $\phi_{num}$ and $\phi_{ord}$ is shown on Figure 2a.

$$m = \sum_{\boldsymbol{x} \in \boldsymbol{X}} [\Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u)) - \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))] \tag{15}$$

where: $m$ margin; $\boldsymbol{x}$ observation; $hit(\boldsymbol{x}, u)$ $u$-th nearest same class neighbor of observation $\boldsymbol{x}$; $miss(\boldsymbol{x}, u)$ $u$-th nearest different class neighbor of observation $\boldsymbol{x}$;

Equation (15) presents final margin form. Parameter $u$ specifies, which nearest neighbor should be taken into consideration. The increase of this parameter decreases the algorithm's sensitivity for outliers, and increases generalization abilities. It is worth noting that the margin in this form is a function of features weights $\alpha(t_o)$. It is important, because it allows weight optimalization. Similar to SIMBA algorithm the gradient ascend method is used. For randomly chosen observation $\boldsymbol{x}$, all $t_o$ parameters are updated according to formula (16) and (17).

$$dt_o = \frac{\partial \Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u))}{\partial t_o} - \frac{\partial \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))}{\partial t_o} \tag{16}$$

$$t_o = t_o + dt_o \tag{17}$$

$$\frac{\partial \Delta(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_o} = \begin{cases} \dfrac{\partial \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_i} & \text{(a)} \\ \dfrac{\partial \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_j} & \text{(b)} \\ \dfrac{\partial \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_k} & \text{(c)} \end{cases} \tag{18}$$

$$\frac{\partial \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_i} = \phi_{num}(x_{1i}, x_{2i}) \left[ \frac{\alpha(t_i) \cdot \phi_{num}(x_{1i}, x_{2i})}{\Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)} \right]^{p-1} \frac{\partial \alpha(t_i)}{\partial t_i} \tag{19}$$

$$\frac{\partial \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_j} = \phi_{cat}(x_{1j}, x_{2j}) \frac{\partial \alpha(t_j)}{\partial t_j} \tag{20}$$

$$\frac{\partial \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_k} = \phi_{ord}(x_{1k}, x_{2k}) \frac{\partial \alpha(t_k)}{\partial t_k} \tag{21}$$

$$\frac{\partial \alpha(t_o)}{\partial t_o} = \frac{1}{\pi(1 + t_o^2)} \tag{22}$$

where: $dt_o$ adjustment for $o$-th feature

Appropriate partial derivative (18) should be chosen depending on the type feature: numerical (a), categorical (b) or ordinal (c). Proper formulas for derivatives have been given in equations (19)–(20). Equation (22) defines a derivative of the chosen weight function depending on the free hidden parameter.

Presented margin form draws very interesting conclusions. Equation (22) shows, that for increasing absolute values $t_o$ algorithm adjustments are declining resulting in stabilization of weights $\alpha(t_o)$ near extreme values 0 or 1 while preserving numerical stability of the solution. Moreover, given the algorithm includes, as special cases, the following algorithms:

1. For $u = 1$, $p = 1$ special case of ReliefF algorithm is get
2. For $\varepsilon_{o1} = 0$, $\varepsilon_{o2} = 1$, $u = 1$, $p = 1$ Relief algorithm is get
3. For $\varepsilon_{o1} = 0$, $\varepsilon_{o2} = 1$, $u = 1$, $p = 2$ SIMBA algorithm is get

Factor $p$ plays special role and it is a form of features mixing factor. Clearly, by increasing $p$, partial derivative (19) is decreasing. Factor in rectangle bracket (19) powered to $p - 1$ is responsible for redundant features reduction in relation to Relief. Therefore, increasing $p$ to reasonable values should decrease redundancy.

## Additional details: multiclass problems, missing values, nearest neighbor choose and final algorithm form

Relief algorithm [11] can also identify important features in case where data set has more than two classes and in case of incomplete data. The same method can be used directly in the presented algorithm. Slightly modifying the formula (16) SIMBAF algorithm can be adapted for many classes. A modified form of the optimized hidden parameter adjustment shows equation (23).

$$dt_o = \sum_{c \neq c(\boldsymbol{x})} \frac{P(c)}{1 - P(c(\boldsymbol{x}))} \frac{\partial \Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u, c))}{\partial t_o} - \frac{\partial \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))}{\partial t_o} \quad (23)$$

where: $P(c)$ probability of class $c$; $c(\boldsymbol{x})$ class of observation $\boldsymbol{x}$; $miss(\boldsymbol{x}, u, c)$ $u$-th nearest class $c$ neighbor of observation $\boldsymbol{x}$.

Another issue is the nearest neighbor choice. Nearest neighbor selection algorithm is sensitive to the curse of dimensionality. For numeric data it has been shown [12], that it might be better to abandon Euclidean metric in favor of Minkowski metrics, even with a fractional order, which better retains the distinction of data points in high-dimensional space. The distance

measure between points is calculated according to formula (24). It should be noted that the parameter $p'$ in the formula below does not correspond to the parameter $p$ from equation (7). Sometimes $p'$ can be as low as $p^{-1}$.

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[\sum_i \left(\frac{|x_{1i} - x_{2i}|}{\delta_i}\right)^{p'}\right]^{1/p'} + \sum_j \phi_{cat}(x_{1j}, x_{2j}) + \sum_k \frac{|x_{1k} - x_{2k}|}{\delta_k} \quad (24)$$

where: $d(*,*)$ distance between two observations; $p'$ metric order.

In case of missing data the following rules were used:

1. When for the first and second observation for a given feature both values are missing, distance and $t_o$ parameter adjustment is set to 0
2. When numerical value is missing, it is replaced by mean value among this feature values
3. When categorical value is missing, distance is a probability that the category is different than in compared observation
4. When ordinal value is missing, it is replaced by median value among this feature values

Listing 1: SIMBAF Algorithm

```
Input data:
    X;                  /*data*/
    N;                  /*feature count*/
    i_max;              /*iteration count*/
    u;                  /*neibour to analize*/
    eps1[N], eps2[N];   /*cutoff parameters*/
    p, p';              /*metric parameters*/
Output data:
    w[N];               /*final weights*/
Variables:
    i,j;
    x;                  /*observation*/
    dt[N], t[N];
Algorithm:
    for j=1 to N
        t[j]=0;
    for i=1 to i_max
        pick up random obserwation x from X
        for j=1 to N
            calculate dt[j] using formula (23)
```

```
        t[]=t[]+dt[]
        for j=1 to N
            calculate w[j] using formula (10)
    return w[];
```

---

**Empirical results and conclusions**

The algorithm described above was used to analyze the effectiveness of infertility treatment by the IVF ICSI/ET method on the data set obtained at the Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok. Specifically designed for this purpose, the application was used to collect these data [2]. The input data set contained a description of 1445 treatment cycles. Each cycle of treatment was represented by 149 independent features (including 107 numerical features, one ordinal feature and 61 categorical features) and one dependant feature – treatment result (pregnant or not). Of course, because of potential treatment process failure at an earlier stage, there were cases of missing data (particularly in describing the characteristics of the final treatment stages). The lower cut-off parameters ($\varepsilon_{o1}$) for numeric and ordinal data was set to 0.1, while the upper ($\varepsilon_{o2}$) to 0.9. Orders of metrics from the equations (7) and (24) were set accordingly to 2.5 ($p$) and 0.5 ($p'$). As a result of the described algorithm execution on the collected data, the following features set was obtained (in order, starting with the largest weight):

- The type of treatment protocol (protocols types described in [2])
- Mucus during ovulation,
- The type of anesthesia at the puncture of ovarian follicles,
- Fallopian factor as a cause of infertility,
- Pain during ovulation,
- Male factor as a cause of infertility,
- Temperature increase during ovulation,
- The number of embryos transferred in the third day of culture,
- Semen preparation – twice washing,
- Hyperprolactinemia in medical history,
- Polycystic ovary syndrome as a cause of infertility,
- Blood spotting during ovulation,
- Type B sperm motility,
- The number of embryos transferred in the second day of culture,

- Type A sperm motility
- Endometriosis as a cause of infertility,
- Age of patient,
- The number of expanding blastocysts in the fifth day of culture.

In the generated as a result features set are those, which had been identified long ago as having a significant impact on the effectiveness of infertility treatment. The addition of other features that previously did not seem to have a significant effect on the treatment result may have a significant impact on improving the quality of prediction based on the so-constructed set of features. This fact confirms the desirability of data collection based on the greatest possible number of features, even potentially non-essential to the treatment process.

The analysis of the above data confirmed also the fact, that the concept of margin plays an important role in the process of building the artificial intelligence mechanisms. Algorithms, that can effectively reduce the number of features to be analyzed, were able to be built using a relatively simple mathematical mechanism, improving the generalizing properties of many classifiers and certainly, the speed of data analysis.

The issue of weight optimization in the presented algorithm requires further study. There are better known methods of global optimization algorithms than used here, e.g. the gradient ascend. The function of the margin itself is highly nonlinear and the algorithm may have a tendency to be stuck in local maximum. Another issue is the form of the optimized margin function itself. In the presented algorithm, it is simply a modified measure of distance (similarity) between observations. The introduction of another margin function could improve the optimization properties. An interesting issue requiring further research are other functions of hidden parameter usage in the role of feature weight.

The next step after identifying a set of significant features is the construction of classifiers from the resulting data set using various methods available. There are many state–of–the–art classifiers whose performance should be checked and the results compared with each other. The aim of further studies will be the development of a series of more–improved predictive models of the effectiveness of infertility treatment using the IVF ICSI/ET method.

## REFERENCES

[1] Radwan J. (ed.) *Niepłodność i rozród wspomagany.* Termedia, Poznań 2005.

[2] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Wołczyński S. The system of electronic registration of information about patients treated for infertility with the IVF ICSI/ET method. *Studies in Logic, Grammar and Rhetoric*, 17 (30), 2009.

[3] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Szamatowicz J., Wołczyński S. Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych. Ginekologia Polska, 80 (12), 2009.

[4] Schapire R. E. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, ed., *Nonlinear Estimation and Classification.* Springer, 2003.

[5] Breiman L. Bagging predictors *Machine Learning*, 24(2), pp. 123–140, 1996.

[6] Boser B. E., Guyon I., and Vapnik V. A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual Workshopon Computational Learning Theory*, pp. 144–152, ACM Press, 1992.

[7] Jollifee I. T. Principal component analysis. Springer Varlag, 1986.

[8] Kohavi R., John G. Wrapper for feature subset selection. *Artificial Intelligence*, 97, pp. 273–324, 1997.

[9] Kira K., Rendell L. A practical approach to feature selection. *Proceedings 9th International Workshop on Machine Learning* pp. 249–256, 1992.

[10] Gilad-Bachrachy R., Navot A., Tishby N. Margin Based Feature Selection – Theory and Algorithms. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[11] Kononenko I., Estimating attributes: analysis and extensions of Relief. Bergadano F., De Raedt L. ed, *Proceedings European Conference on Machine Learning*, 1994.

[12] Aggarval Ch. C, Hinneburg A., Keim D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Lecture Notes in Computer Science*, 1973/2001, Volume, pp. 420–434, 2001.